Generation of Captions Highlighting the Differences between a Clothing Image Pair with Attribute Prediction

Kohei Abe
Graduate School of Information Science and Technology, Hokkaido University
Sapporo, Hokkaido, Japan
email: ko.abe@ist.hokudai.ac.jp

Soichiro Yokoyama
Faculty of Information Science and Technology, Hokkaido University
Sapporo, Hokkaido, Japan
email: yokoyama@ist.hokudai.ac.jp

Tomohisa Yamashita
Faculty of Information Science and Technology, Hokkaido University
Sapporo, Hokkaido, Japan
email: yamashita@ist.hokudai.ac.jp

Hidenori Kawamura
Faculty of Information Science and Technology, Hokkaido University
Sapporo, Hokkaido, Japan
email: kawamura@ist.hokudai.ac.jp

***Abstract*—Detailed information for comparisons between products is necessary in consumers' product purchasing process, especially during the information search and choice evaluation phases. However, conventional product descriptions, which are the main source of information, tend to focus only on the product in question, and thus do not adequately express the differences between products. To solve this problem, garments are treated as target products, and a caption-generation method that emphasizes the differences between pairs of garment images using a deep-learning model for image caption-generation is proposed and its effectiveness verified. The proposed method selects and outputs captions that express differences in features from a set of captions generated for input-garment image pairs. Subject experiments confirmed that the proposed method accurately represented the feature differences between garments and provided useful information for consumers to compare garments. In particular, the proposed method is highly effective for garment pairs with similar features.**

Keywords-deep learning; image captioning; consumer support; information provision.

I. Introduction

In the field of consumer behavior, the sequence of processes involved in the purchase of a product is widely recognized as the purchase decision-making process [1]. This process comprises five stages: problem recognition, information search, alternative evaluation, purchase decisions, and post-purchase evaluation. In the problem recognition phase, consumers identify their needs and problems, and collect information to satisfy them in the information search phase. In the evaluation of alternatives, the consumer compares and evaluates products based on the collected information, and selects and purchases a specific product in the purchase decision stage. In the post-purchase evaluation, the degree of satisfaction was determined based on the results of the product use. During the information search and evaluation of alternatives phase, consumers need detailed information to understand the characteristics and differences of products and make the right choices. This information can originate from a variety of sources, such as user reviews, expert opinions, and comparison websites; however, product descriptions are one of the most important sources of information that consumers interact with in the early stages of their purchasing decisions. Product descriptions can successfully convey the basic features of a product; however, they tend to focus only on the product in question and do not adequately describe the differences between products. This lack of information may affect consumers' final purchasing decisions and post-purchase evaluations.

Image-caption generation is a research area for generating descriptive text from images; however, it primarily generates a single sentence for a single input image. It is impossible to generate a caption for each image by considering the relationships between multiple images. Some studies have aimed to generate distinctive image captions by comparing input images with similar images in a database; however, they cannot specify the images to be compared, as was the aim of this study.

This study aimed to provide adequate information to consumers when comparing products. As a concrete initial effort towards this goal, a method for generating captions that highlight the differences between two products is proposed and evaluated. Clothing is selected as the target product. Clothing is an everyday purchase for consumers and has various features, such as pattern, material, length, and collar shape. Therefore, consumers need to compare product features during product selection. In the proposed method, two different garment images are independently input into an image-caption generation model to generate multiple captions. Next, the prominence of each attribute in each image is calculated using the garment attribute estimation model and the frequency of occurrence in the caption. This is compared between images, and the caption containing more salient attributes than one image is selected from the multiple captions generated for each

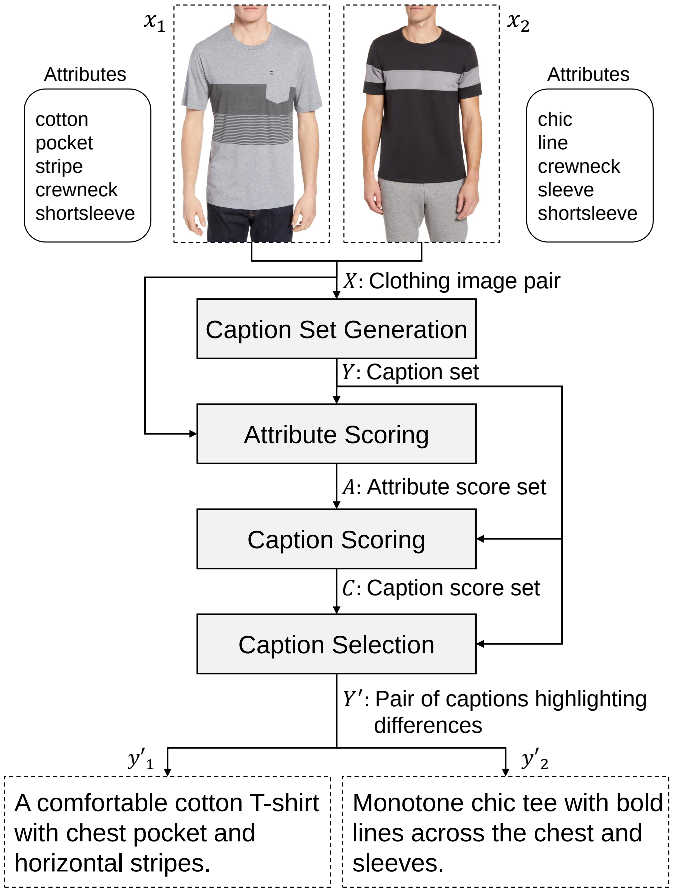


Figure 1. Overview of the proposed method.

image and output. The proposed caption-generation method yields captions that contain more salient features than one garment, with one sentence for each image and an average of approximately 14 words. Examples of the captions obtained are shown in Figure 1. In the subject experiment, it was evaluated whether the captions obtained using the proposed method contained obvious errors, how well they described features that were only present in one garment, and whether they were useful for comparing garments. This experiment confirmed that the generated captions adequately described the differences between products and provided useful information for product comparison.

The remainder of this paper is organized as follows. Section II describes work related to this study. Section III describes the proposed method. Section IV describes in detail the models and datasets used in the experiments. Section V describes the experiments on the comparative validation of the proposed method by employing different scoring methods. Section VI describes the experiments that qualitatively evaluate the captions generated by the proposed method. Finally, Section VII discusses the conclusions of this study and future perspectives.

TABLE I
COMPARISON OF IMAGE CAPTION GENERATION MODELS

Model	BLEU4	METEOR
NIC [5]	27.7	23.7
NICA [8]	25.0	13.9
SCST [9]	31.9	25.5
ClipCap [10]	33.5	27.5
OFA [13]	44.9	32.5

II. RELATED WORK

This section describes the main areas relevant to this study, namely image caption generation, caption generation for multiple images, garment attribute estimation, and garment image caption generation.

A. Image Caption Generation

Image-caption generation is the task of generating an appropriate description of a single-input image. A comparison of the main image-caption generation models for the benchmark dataset Microsoft Common Objects in Context (MS COCO) [2] is presented in Table I. Bilingual Evaluation Understudy (BLEU) [3] and Metric for Evaluation of Translation with Explicit Ordering (METEOR) [4] are automatic metrics that measure the similarity between the generated and correct captions, with higher values indicating better model performance. Vinyals et al. [5] proposed a model based on a deep recurrent architecture that combines a Convolutional Neural Network (CNN) [6] and Long Short Term Memory (LSTM) [7]. Subsequently, Xu et al. [8] introduced an attention mechanism that focused on specific regions in an image when generating different words. Furthermore, Rennie et al. [9] proposed a model that incorporates reinforcement learning. Recently, image-language pre-training models that learn using large amounts of image-text pair data have achieved higher accuracy than conventional models. Mokady et al. [10] proposed a model that combines the image language pre-training model Contrastive Language–Image Pre-training (CLIP) [11] and the language model Generative Pre-trained Transformer 2 (GPT-2) [12], which reduces training time and achieves highly accurate caption generation. Wang et al. [13] also proposed a pre-training model using 20 million image-text pair data. All these models generate a single-sentence caption for a single input image. In this study, one-sentence captions are generated for each of the two input images. A one-input, one-output image caption generation model is used independently to generate multiple captions for each input image. Each caption is then scored, and the highest caption is generated one sentence at a time to generate a one-sentence caption for each of the two images.

B. Caption Generation for Multiple Images

Several efforts have been made to generate captions for multiple images as an application of conventional image-caption generation. One example is the change in the image-caption generation initiative. This method identifies changes between two input images and generates a one-sentence caption describing the change [14][15]. In this study, a caption is generated for each input image. In conventional image-caption generation, which tends to generate generic sentences, the distinctive parts of the input images are often ignored. To address this problem, an approach called feature-based image-caption generation is currently in progress [16][17]. In this approach, a single input image is compared to a set of similar images in a database to identify the distinctive aspects of the input image, which are then reflected in the caption. However, this approach does not specify similar images explicitly. In this study, two specified images are compared. The attribute estimates calculated for each image are compared, and a relative score is calculated. The caption score is then calculated by summing the attribute estimates that appear in the caption and is used for caption selection.

C. Clothing Attribute Estimation

Clothing attribute estimation is the task of estimating features, such as the material, pattern, collar shape, and sleeve length of clothing in an image. Examples of the estimated attributes include cotton, floral, sleeveless, and leather. This task has been applied to garment retrieval and recommendation. Chen et al. [23] proposed a model that combines a CNN [25] trained on a large image dataset, ImageNet [24] with a multilayer perceptron for a garment image retrieval task that matches images of garments worn by a person with those from a fashion e-commerce site. Similarly, Huang et al. [26] proposed a deep model that included two CNNs to handle street images and e-commerce site images in garment image retrieval. Both models were trained using bounding boxes to identify garment regions. In contrast, Liu et al. [21] proposed a model that learns garment landmark information, such as sleeve and collar positions, estimates the landmarks during inference, and uses this information as an aid for garment attribute estimation. A comparison of the garment-attribute estimation models on the benchmark dataset, Deepfashion [21] is presented in Table II. The Top-k Recall [22] was used as an evaluation metric. This assigns the top-k attributes with the highest probability of estimation to each image and measures the number of correctly estimated attributes. By estimating landmark information, FashionNet can better recognize the shape and position of garments and perform better than models that use only bounding boxes. Here, consumer perceptions of attributes are subjective and depend on age and gender. Different consumers may consider different attributes important when comparing garments. However, as a first attempt in this study, the weighting of the attributes did not change. Only estimates objectively calculated using the model were used.

D. Clothing Image Caption Generation

Sonoda et al. [18] proposed a method for searching for similar input images from a set of garment images they collected and applied the obtained garment information and features of similar images to a template. Yang et al. [19] proposed a framework that supports the creation of product introductions on e-commerce websites. In their study, attribute- and sentence-level rewards were introduced to improve the quality of captions generated. They also adopted a method for integrating the training of the model using maximum likelihood estimation, attribute embedding, and reinforcement learning. In addition, a large dataset for garment image-caption generation containing approximately one million images was constructed. Cai et al. [20] removed noisy garment images and reconstructed a clean garment image dataset. These studies generated captions describing the salient features of a single-input garment image. They are insufficient for the purpose of this research, that is, to provide information when comparing garments, in that they cannot express the detailed differences between different garments. In this study, a caption is generated that highlights the differences between two input garment images.

TABLE II
COMPARISON OF CLOTHING ATTRIBUTE ESTIMATION MODELS

Model	Top-3 Recall	Top-5 Recall
WBIT [23]	27.46	35.37
DARN [26]	40.35	50.55
FashionNet [21]	45.52	54.61

III. Proposed Method

This section describes the caption-generation method proposed in this study, which highlights the differences between garment image pairs. An overview of the method is presented in Figure 1. The method considers a pair $X = \{x_i \mid i = 1, 2\}$ of different garment images as input and outputs a caption pair $Y' = \{y'_i \mid i = 1, 2\}$ corresponding to each image, where x_i is the i-th garment image, and y'_i is the output caption corresponding to x_i. In Figure 1, the attribute set annotated to the image is displayed next to each image. This method comprises four modules: caption set generation, attribute scoring, caption scoring, and caption selection. The following sections describe these modules in detail.

A. Caption Set Generation Module

The caption set generation module considers a pair X of different garment images as input, inputs each image independently of the image caption-generation model, and outputs a caption set $Y = \{y_{ij} \mid i = 1, 2; j = 1, 2, \ldots, J\}$ corresponding to each image. Here, y_{ij} represents the j-th caption for image x_i. The image-caption generation model used in this study is described in detail in Section IV.

B. Attribute Scoring Module

The attribute scoring module considers a pair of different garment images X and a caption set Y as input and outputs a set of attribute scores $A = \{a_{ik} \mid i = 1, 2; k \in K\}$ for each image. Here, K is the set of attributes to be evaluated and a_{ik} is the score of attribute k for image x_i. An attribute score is a numerical expression of the prominence of a particular attribute exhibited by a garment image; the higher the score, the stronger the garment image that exhibits that attribute. An example of an attribute score for the garment image x_1 in Figure 1 is 0.20 for crewneck, 0.15 for pocket, and 0.01 for sleeveless, which were calculated to be higher when the image had the attribute prominently and lower when it did not. In this study, two methods of attribute scoring were considered: attribute scoring based on attribute estimation, and attribute scoring based on frequency of occurrence.

1) Attribute Scoring Based on Attribute Estimation: Attribute scoring based on attribute estimation uses a garment-attribute estimation model, whose output is the estimated probability of each attribute for an input-garment image. The estimated probability of an attribute for each image was calculated, and this value was used as the attribute score. This is illustrated in (1), where p_{ik} is the estimated probability of attribute k for image x_i. The clothing attribute estimation model used in this study is described in detail in Section IV.

$$a_{ik} = p_{ik} \tag{1}$$

2) Attribute Scoring Based on Frequency of Occurrence: The caption generated for each garment image using the caption set generation module reflects the garment characteristics. If a particular attribute appears frequently in a caption set, it can be regarded as one of the main features of the garment. This method calculates the frequency of occurrence of each attribute in the caption set for each image and uses this value as the attribute score. This is illustrated in (2), where f_{ijk} is the number of occurrences of attribute k in the caption y_{ij}.

$$a_{ik} = \frac{1}{J}\sum_{j=1}^{J} f_{ijk} \tag{2}$$

C. Caption Scoring Module

The caption scoring module considers a caption set Y and an attribute score set A as inputs, and outputs a caption score set $C = \{c_{ij} \mid i = 1, 2; j = 1, 2, \ldots, J\}$. The caption score is a numerical expression of the extent to which the caption reflects the salient attribute differences between the garment images and attributes specific to each image; a higher score is regarded as emphasizing the differences between one image and the other. Here, c_{ij} represents the score of the caption y_{ij}. In this study, two caption scoring methods were considered: caption scoring based on the comparison of top attributes and caption scoring based on the addition of relative scores. These methods are described in detail as follows.

1) Caption Scoring Based on Comparison of Top Attributes: This method first obtains an attribute set K_i^{top-n} with the top n attribute scores for each image. Next, the difference set D_i of K_i^{top-n} for each image is the difference attribute set, and the product set T is the common attribute set. These are presented in (3)∼(5):

$$D_1 = K_1^{top-n} \setminus K_2^{top-n} \tag{3}$$

$$D_2 = K_2^{top-n} \setminus K_1^{top-n} \tag{4}$$

$$T = K_1^{top-n} \cap K_2^{top-n} \tag{5}$$

Finally, the difference between the number of attribute occurrences in the different attribute sets and the number of attribute occurrences in the common attribute set for each caption was calculated and used as a caption score. This process is illustrated in (6), where f_{ijk} is the number of occurrences of attribute k in caption y_{ij}.

$$c_{ij} = \sum_{k \in D_i} f_{ijk} - \sum_{k \in T} f_{ijk} \tag{6}$$

This method assigns higher scores to captions containing more differentiated and fewer common attributes.

2) Caption Scoring Based on Relative Score Addition: This method first calculates the difference in attribute scores between images to obtain the relative attribute scores Δa_{ik}. These are given by Equations (7) and (8), respectively.

$$\Delta a_{1k} = a_{1k} - a_{2k} \tag{7}$$

$$\Delta a_{2k} = a_{2k} - a_{1k} \tag{8}$$

Next, the relative attribute scores corresponding to the attributes in the caption are added and used as the caption score. The process is described in (9), where $K_{y_{ij}}$ represents the set of attributes contained in the caption y_{ij}.

$$c_{ij} = \sum_{k \in K_{y_{ij}}} \Delta a_{ik} \tag{9}$$

Using this method, captions containing more attributes with relatively high attribute scores have higher scores.

D. Caption Selection Module

The caption selection module considers the caption sets Y and C as input, selects the caption with the highest caption score in the caption set corresponding to each image, and outputs a set of captions $Y' = \{y'_i \mid i = 1, 2\}$ that highlights the differences. This process is represented by (10).

$$y'_i = \underset{y_{ij}}{\mathrm{argmax}} c_{ij} \tag{10}$$

IV. Models and Datasets

This section describes the image-caption generation models, garment attribute estimation models, and garment image datasets used in the study.

A. Image Caption Model and Clothing Attribute Estimation Model

This study is looking at reflecting different national and regional fashion cultures in captions in the future. Therefore, image-caption generation models that can handle garment image data in various languages are desirable. Among the image-caption generation models compared in Section II, ClipCap [10] is a combination of CLIP and the language model GPT-2. It is easy to handle non-English data because CLIP exists for multiple languages [29], and the language model Generative Pre-trained Transformer 4 (GPT-4) [30], which is similar to GPT-2, supports multiple languages. Furthermore, as shown in Table I, the accuracy is sufficiently high among the major image-caption generation models. Therefore, in this

TABLE III
COMPARISON OF CLOTHING IMAGE DATASETS

Dataset	Number of images	Attributes	Captions
FACAD170K [20]	178,849	yes	yes
DeepFashion [21]	289,222	yes	no
FashionGen [27]	325,536	no	yes
iFashion [28]	1,062,550	yes	no

study, ClipCap was used as the image-caption generation model in the caption set generation module. FashionNet [21] was used as the garment-attribute estimation model in the attribute-scoring module. This model estimates the landmarks of a garment and uses the obtained information for garment attribute estimation. This model can capture the fine-grained features of a garment image and is highly accurate.

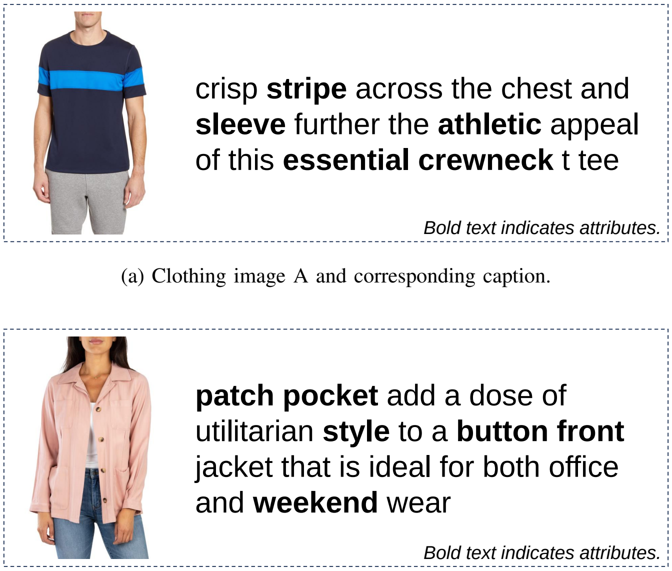
B. Clothing Image Dataset

A comparison of the main garment image datasets is shown in Table III. In this study, the FACAD170K garment image dataset [20] with both attributes and captions, which enables an attribute-based caption evaluation, was used to train the image-caption generation model. An example of the FACAD170K data is shown in Figure 2. Each garment image was crawled from a generic website, mainly Google Chrome, and was either an image of a person wearing the garment or an image of the garment alone, with a one-sentence caption from the web. The data collected using this method reflect the variety of styles and trends in clothing that real consumers interact with on a daily basis and are therefore highly suitable for simulation and analysis to mimic the context of consumers' clothing choices. The same caption is provided for garments of different colors. The bold text in the captions for Figure 2 represents multiple attributes assigned to a single garment image. FACAD170K has 990 attributes. In contrast, training the garment-attribute estimation model requires bounding boxes and landmark information to identify garment regions. However, FACAD170K did not contain these annotations. Because annotation is time-consuming, we used FashionNet's Deepfashion [21] pretraining model for garment attribute estimation. DeepFashion contains 1000 attributes, 292 of which match FACAD170K. The top ten attributes with the highest frequency of occurrence in FACAD170K and their frequencies are listed in Table IV. FACAD170K and DeepFashioin data with these attributes were used to evaluate the proposed method.

V. COMPARATIVE VERIFICATION OF ATTRIBUTE AND CAPTION SCORING METHODS

A. Objectives

This experiment aimed to compare and validate attribute scoring based on attribute estimation and frequency of occurrence in the attribute scoring module and caption scoring based on the comparison of top attributes and the addition of relative scores in the caption scoring module to find the best combination of methods for generating captions that highlight differences.


crisp **stripe** across the chest and **sleeve** further the **athletic** appeal of this **essential crewneck** t tee

Bold text indicates attributes.

(a) Clothing image A and corresponding caption.

patch pocket add a dose of utilitarian **style** to a **button front** jacket that is ideal for both office and **weekend** wear

Bold text indicates attributes.

(b) Clothing image B and corresponding caption.

Figure 2. Examples of data from the FACAD170K dataset.

TABLE IV
HIGH-FREQUENCY ATTRIBUTES COMMON TO BOTH FACAD170K AND DEEPFASHION

Attribute	Frequency (%)
cotton	4.53
cut	4.41
soft	3.76
sleeve	2.98
fit	2.81
leather	2.58
stretch	2.46
classic	2.45
knit	2.31
strap	2.25

B. Methods

In this experiment, the captions generated using the four proposed methods were automatically evaluated. In the caption set generation module, the image-caption generation model ClipCap was trained using 177,849 training data points from FACAD170K. The key parameters during training were set to a learning rate of 2.0×10^{-5}, a batch size of 40, and 10 epochs. These parameters were set based on the settings used in the original study [10]. $J = 100$ captions were generated for each image, based on the probability distribution of the language model. In the attribute scoring module, 292 attributes common to FACAD170K and DeepFashion were used as the attribute set K to be evaluated. Caption scoring based on top attribute comparisons in the caption scoring module uses the top $n = 9$ attributes. The values were determined based on preliminary experiments that compared the estimated and correct attributes for different values of n. The model was evaluated by comparing the inferred results of the model against FACAD170K and DeepFashion with correct labels. The evaluation metrics are as follows. The set of attributes annotated for a garment image x_i is the overall attribute set K_i^{GT}, and the set of attributes with only one garment image is the differential attribute set D_i^{GT}. This is expressed in (11) and (12).

$$D_1^{GT} = K_1^{GT} \setminus K_2^{GT} \tag{11}$$

$$D_2^{GT} = K_2^{GT} \setminus K_1^{GT} \tag{12}$$

Let $K_{y_i'}$ be the attribute set contained in the caption y_i'. The precision, Recall, and F1 scores were calculated between $K_{y_i'}$ and the differential attribute set D_i^{GT} to assess the degree of description of the attributes that differed between garments. Similar indices were calculated between $K_{y_i'}$ and the overall attribute set K_i^{GT} as supplementary indices to assess the degree of description of the attributes in each garment image. Larger values of these indices are preferable. The evaluation was performed on 10,000 pairs, and the average value of each evaluation indicator was calculated.

C. Results

The evaluation results for the captions generated by the proposed method in FACAD170K and DeepFashion are listed in Tables V and VI. A comparison of the results across datasets shows that the evaluation values for FACAD170K are higher than those of DeepFashion for all indicators. This is because the image-caption generation model ClipCap was trained on the FACAD170K data; consequently, the attribute information of FACAD170K was more appropriately reflected in the captions. For attribute scoring methods, frequency-of-occurrence-based attribute scoring tends to perform better than attribute estimation-based attribute scoring on both datasets. In particular, FACAD170K outperformed the attribute scoring based on attribute estimation for all evaluation indicators. Regarding caption scoring methods, caption scoring based on relative score addition outperformed caption scoring based on top-attribute comparisons for all evaluation indices in both datasets. These results indicate that under the experimental conditions of this study, the combination of attribute scoring based on the frequency of occurrence and caption scoring based on relative score addition is the most effective.

VI. Qualitative Evaluation of Generated Captions

A. Objectives

This experiment aimed to assess how accurately the captions generated by the proposed method represent the features of a single garment, how well they capture the differences between garment image pairs, and how useful they are for comparing garments.

B. Methods

In this experiment, the captions generated using the proposed method were presented to a group of subjects for evaluation. The subject group comprised ten male and female subjects in their 20s. Clothing image pairs and captions are shown in Figure 3. Five pairs of clothing images were prepared (Pairs 1 to 5). To compare the effectiveness of the proposed method based on the similarity between garments, Pairs 1 to 3 have high similarity, whereas Pairs 4 and 5 have low similarity. They were selected based on visual confirmation and the degree of agreement between the attributes of each garment image. The individual garment images were assigned a name, such as 1A for the image on the left side of Pair 1 and 1B for that on the right side. The proposed method employs the method that achieved the best performance in the experiments described in Section V. Specifically, attribute scoring based on frequency of occurrence and caption scoring based on relative score addition were applied. The questions and options set are shown in Table VII. Q1 was designed to assess how accurately the caption represented garment characteristics. Q2 and Q3 assessed how well the captions described the features of only one garment. Furthermore, Q4 was established to test the usefulness of the caption pairs provided for comparing garments. A five-point Likert scale was used to answer each question. In addition, the subjects were asked to explain the reasons for their choice of options and any erroneous features and features not described in the caption. Wilcoxon's signed-rank test was used as the test method. This test checked whether the answers to each question were significantly biased from neutral and the significance level was set at 5%. A similar test was used to examine the difference in responses between the two questions. A significant difference between the distribution of responses to the two questions was tested to determine whether a significant difference existed. For the comparative analysis based on the similarity between clothing image pairs, the Mann-Whitney U test was employed because of the different sample sizes, and the significance level was set at 5%. This analysis examined whether significant differences existed in the distribution of responses between garment pairs with different similarities. Furthermore, Bonferroni correction was applied to account for the effects of multiple tests.

C. Results

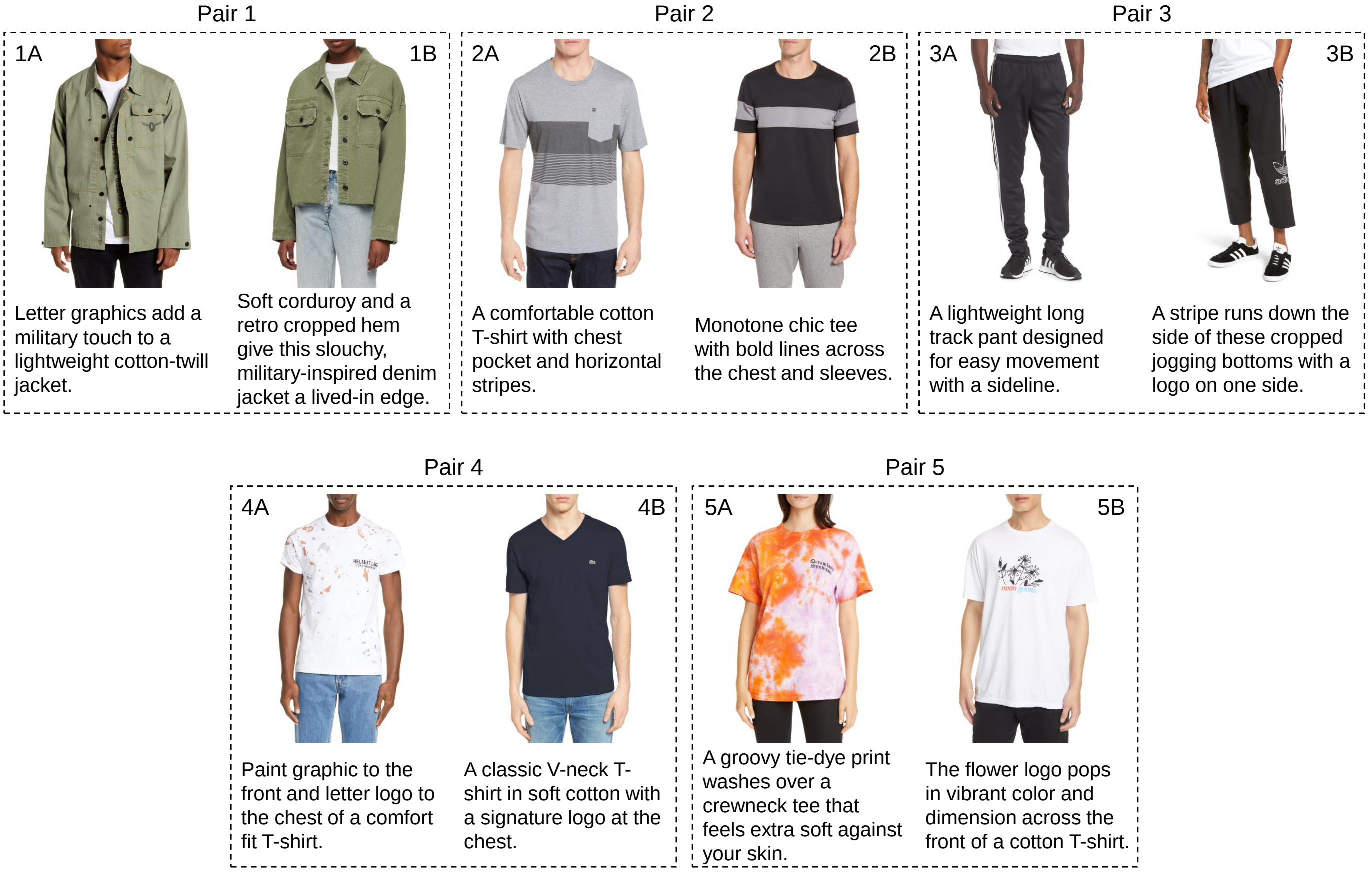
The proportions of the responses to each question are shown in Figure 4. In Q1, approximately 60% of the respondents answered 'strongly disagree' and the p-value of the Wilcoxon signed-rank test was 4.02×10^{-12}, indicating a bias towards negative opinions rather than neutrality. For Q2, all responses were 'strongly agree' or 'agree', with a Wilcoxon test p-value of 6.50×10^{-22}, indicating a bias towards positive opinions rather than neutrality. In Q3, the proportion of respondents who answered 'strongly agree' was approximately 50% lower than that in Q2, but the p-value of the Wilcoxon test was 1.25×10^{-13}, indicating a bias towards positive rather than neutral opinions. The p-value of the Wilcoxon test between the responses to Q2 and Q3 is 1.27×10^{-8}, confirming a significant difference between the two questions. In Q4, the total number of 'strongly agree' and 'agree' responses reached approximately 80%, with a Wilcoxon test p-value of 5.41×10^{-5}, indicating a bias towards more positive than neutral opinions. The percentages of responses to Q4 in clothing image pairs with high and low similarity are shown in Figure 5. The Mann-Whitney U-test results showed a p-value of 1.03×10^{-3}, confirming a significant difference.

TABLE V
RESULTS IN FACAD170K

Attribute Scoring	Caption Scoring	Differential Attributes			Overall Attributes		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
Attribute Estimation	Comparison of Top Attributes	0.145	0.171	0.144	0.198	0.174	0.173
	Relative Score Addition	0.157	0.223	0.172	0.212	0.225	0.206
Frequency of Occurrence	Comparison of Top Attributes	0.204	0.324	0.236	0.248	0.294	0.258
	Relative Score Addition	0.214	0.369	0.256	0.274	0.353	0.297

TABLE VI
RESULTS IN DEEPFASHION

Attribute Scoring	Caption Scoring	Differential Attributes			Overall Attributes		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
Attribute Estimation	Comparison of Top Attributes	0.051	0.089	0.058	0.077	0.091	0.077
	Relative Score Addition	0.070	0.136	0.084	0.123	0.164	0.131
Frequency of Occurrence	Comparison of Top Attributes	0.057	0.139	0.075	0.088	0.143	0.104
	Relative Score Addition	0.059	0.156	0.080	0.096	0.171	0.118



1A: Letter graphics add a military touch to a lightweight cotton-twill jacket.

1B: Soft corduroy and a retro cropped hem give this slouchy, military-inspired denim jacket a lived-in edge.

2A: A comfortable cotton T-shirt with chest pocket and horizontal stripes.

2B: Monotone chic tee with bold lines across the chest and sleeves.

3A: A lightweight long track pant designed for easy movement with a sideline.

3B: A stripe runs down the side of these cropped jogging bottoms with a logo on one side.

4A: Paint graphic to the front and letter logo to the chest of a comfort fit T-shirt.

4B: A classic V-neck T-shirt in soft cotton with a signature logo at the chest.

5A: A groovy tie-dye print washes over a crewneck tee that feels extra soft against your skin.

5B: The flower logo pops in vibrant color and dimension across the front of a cotton T-shirt.

Figure 3. Presented clothing image pairs and captions shown to participants.

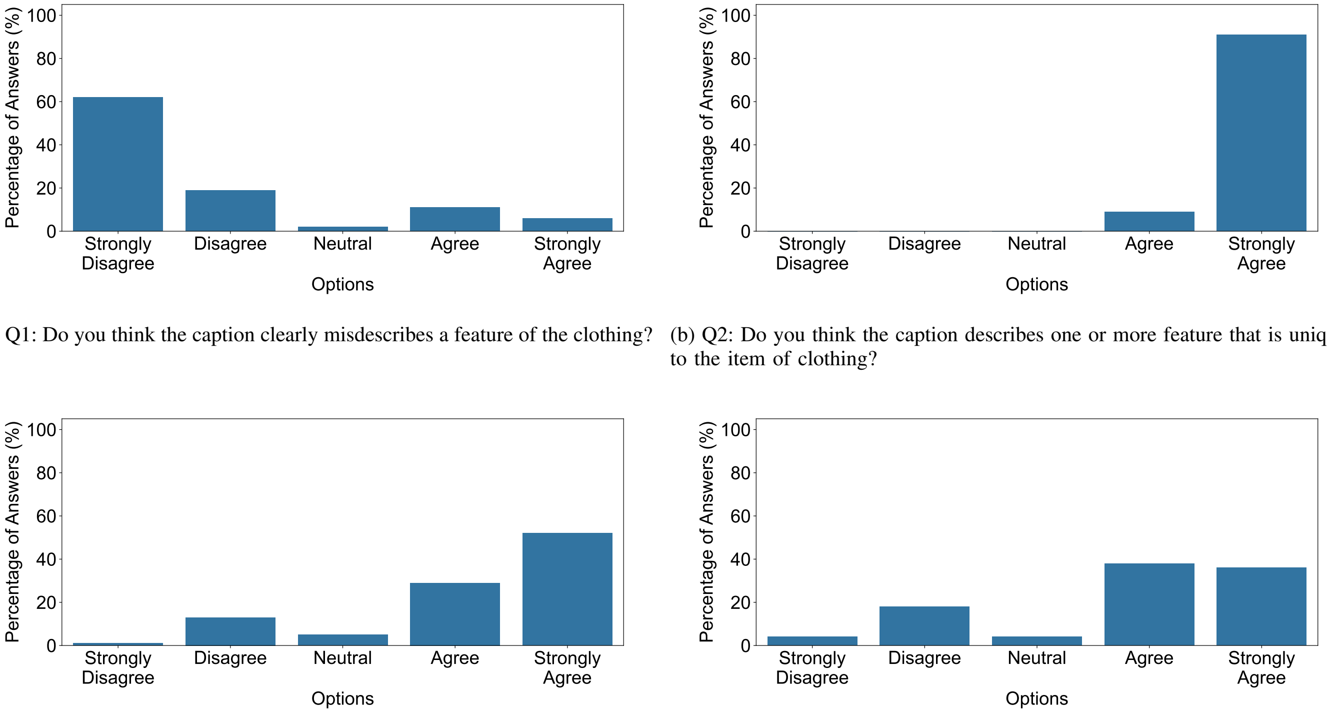
D. Discussions

The results for Q1 suggest that the captions generated by the proposed method accurately describe the characteristics of the garment. However, several users pointed out that garment image 1B, which is made of denim fabric, was incorrectly described as a corduroy material. The corduroy attribute appeared nine times in the caption set for garment image 1B, compared to zero times for garment image 1A, resulting in a higher attribute score, and captions containing the corduroy attribute were preferentially selected. This can be attributed to the difficulty in recognizing detailed materials using CLIP. As some subjects judged this difference in material to be non-erroneous, this feature is also difficult for humans to identify.

The results for Q2 and Q3 suggest that the captions may

TABLE VII
SET QUESTIONS AND OPTIONS

	Question	Options
Q1	Do you think the caption clearly misdescribes a feature of the clothing?	• Strongly Agree • Agree • Neutral • Disagree • Strongly Disagree
Q2	Do you think the caption describes one or more feature that is unique to the item of clothing?	
Q3	Do you think the caption describes all the features that are unique to the item of clothing?	
Q4	Do you think that the two captions would help you to compare the clothing if you were deciding whether to buy one of the clothing items?	



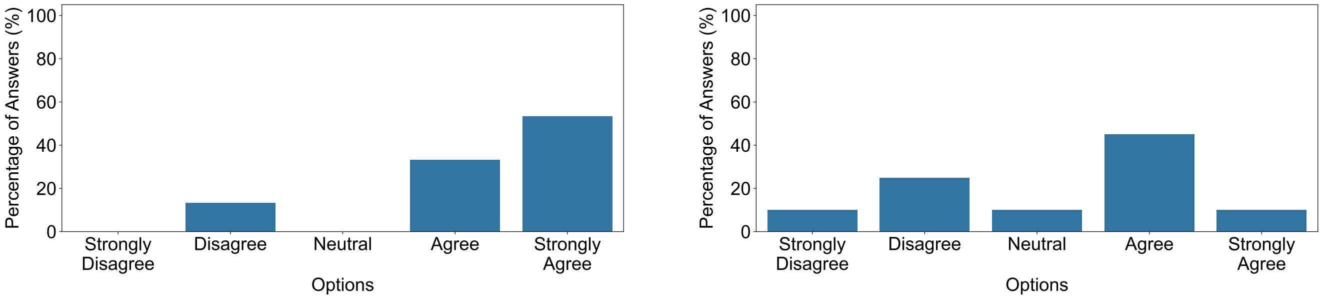
(a) Q1: Do you think the caption clearly misdescribes a feature of the clothing?

(b) Q2: Do you think the caption describes one or more feature that is unique to the item of clothing?

(c) Q3: Do you think the caption describes all the features that are unique to the item of clothing?

(d) Q4: Do you think that the two captions would help you to compare the clothing if you were deciding whether to buy one of the clothing items?

Figure 4. Percentage of answers to each question.



(a) Similar pairs.

(b) Dissimilar pairs.

Figure 5. Percentage of answers to Q4 for similar and dissimilar pairs.

describe at least one feature unique to a garment, but not all of them exhaustively. An example of an exhaustive description is provided in Pair 3. Differences in length and the presence or absence of the logo were described in the caption, and many respondents indicated that all features unique to one garment were described in the caption. Conversely, as examples of non-exhaustive descriptions, it was pointed out that Pair 1 did not describe the number of pockets and Pair 2 did not describe the different colors of the garments. The attribute 'pocket' allows the presence or absence of pockets to be reflected in the caption, but it is considered difficult to reflect the number of pockets. Regarding color, although attributes for color exist in FACAD170K, the same caption is given to garment images of different colors, and there are few descriptions of color, which may be attributed to the difficulty in generating descriptions for color.

The results for Q4 showed that the generated captions provided useful information for the comparison of garments. Furthermore, they were more useful for pairs with high similarity than for those with low similarity. Some participants commented that reading the captions helped them focus on features of highly similar pairs that were not immediately noticeable in the images, such as the differences in length and graphics for Pair 1, position of the lines for Pair 2, and differences in length for Pair 3. However, in the less similar pairs, Pairs 4 and 5, the differences in the features pointed out by the captions were visually clear, and many were critical to the usefulness of the captions in the comparison. As there were no opinions that the captions described differences in features that were difficult to notice, it was considered that captions highlighting differences in pairs with low similarity were not useful.

VII. Conclusion and Future Work

In this study, a caption-generation method that highlights the differences between pairs of garment images to provide useful information for consumers when comparing products was proposed and evaluated. In this method, two different garment images are first input independently into an image caption generator to generate multiple captions. Attribute scores are then calculated for each image. A caption score is then calculated for each caption in the multiple captions generated for each image using the attribute scores. Finally, the captions are selected and output based on caption scores. Automatic evaluation experiments were conducted on attribute scoring and caption scoring, focusing on accurately describing the features of a single garment and the differences between garments. Methods employing attribute scoring based on the frequency of occurrence and caption scoring based on relative score addition were rated highly. Attribute scoring based on frequency of occurrence uses the frequency of an attribute's occurrence in the caption as the attribute score, whereas caption scoring based on relative score addition calculates the relative value of the attribute score and adds it to the number of attributes that appear. Furthermore, captions generated by a combination of methods that received high ratings in the automatic evaluation experiment were presented to the subjects, and a qualitative evaluation of their usefulness was conducted. The results confirm that the proposed method provides useful information for comparing two garments. It was also confirmed that the proposed method is more effective for highly similar garment pairs than for less similar garment pairs. As it is assumed that consumers often compare garments with high similarity when comparing garments, an approach for garments with high similarity is planned.

The proposed method can only specify two garment images as input images. We plan to extend this approach to handle more than three garment images to better meet consumer garment comparison needs. Specifically, we believe that the relative scores can be calculated in the same manner as in the present study by subtracting the average attribute scores of the other images from the attribute score of one garment during attribute scoring. In addition, the performance of the proposed method is highly dependent on the accuracy of the image caption generation model and the diversity of the generated captions. Because the image caption generation model can be easily changed to other models because of the structure of the proposed method, there is room to verify its performance when using state-of-the-art models, such as GPT-4 Vision.

Acknowledgment

We would like to thank Editage (www.editage.jp) for English language editing.

References

[1] M. R. Solomon, "Consumer behavior : buying, having, and being," 12th ed, Pearson Education, 2016.

[2] T.-Y. Lin et al., "Microsoft coco: Common objects in context," Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pp. 740-755.

[3] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311-318, 2002.

[4] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp. 65-72, 2005.

[5] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156-3164, 2015.

[6] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," International conference on machine learning, pp. 448-456, 2015.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[8] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," International conference on machine learning, pp. 2048-2057, 2015.

[9] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7008-7024, 2017.

[10] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," arXiv preprint arXiv:2111.09734, 2021.

[11] A. Radford et al., "Learning transferable visual models from natural language supervision," International conference on machine learning, pp. 8748-8763, 2021.

[12] A. Radford et al., "Language models are unsupervised multitask learners," OpenAI blog, vol. 1, no. 8, pp. 9, 2019.

[13] P. Wang et al., "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," International Conference on Machine Learning, pp. 23318-23340, 2022.

[14] H. Jhamtani and T. Berg-Kirkpatrick, “Learning to describe differences between pairs of similar images,” Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4024-4034, 2018.

[15] D. H. Park, T. Darrell, and A. Rohrbach, “Robust change captioning,” Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4624-4633, 2019.

[16] J. Wang, W. Xu, Q. Wang, and A. B. Chan, “Group-based distinctive image captioning with memory attention,” Proceedings of the 29th ACM International Conference on Multimedia, pp. 5020-5028, 2021.

[17] Y. Mao et al., “Rethinking the reference-based distinctive image captioning,” Proceedings of the 30th ACM International Conference on Multimedia, pp. 4374-4384, 2022.

[18] A. Sonoda and G. Niina, “Apparel EC saito ni okeru setsumei bun jidou seisei (Automatic Generation of Descriptions in Apparel E-commerce Sites),” Proceedings of the Japan Society of Management Information National Conference, pp. 125-127, 2018.

[19] X. Yang et al., “Fashion captioning: Towards generating accurate descriptions with semantic rewards,” Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16, pp. 1-17.

[20] C. Cai, K.-H. Yap, and S. Wang, “Attribute conditioned fashion image captioning,” IEEE International Conference on Image Processing, pp. 1921-1925, 2022.

[21] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, “Deepfashion: Powering robust clothes recognition and retrieval with rich annotations,” Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1096-1104, 2016.

[22] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, “Deep convolutional ranking for multilabel image annotation,” arXiv preprint arXiv:1312.4894, 2013.

[23] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, “Where to buy it: Matching street clothing photos in online shops,” Proceedings of the IEEE international conference on computer vision, pp. 3343-3351, 2015.

[24] J. Deng et al., “ImageNet: A Large-Scale Hierarchical Image Database,” IEEE conference on computer vision and pattern recognition, pp. 248-255, 2009.

[25] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[26] J. Huang, R. S. Feris, Q. Chen, and S. Yan, “Cross-domain image retrieval with a dual attribute-aware ranking network,” Proceedings of the IEEE international conference on computer vision, pp. 1062-1070, 2015.

[27] N. Rostamzadeh et al., “Fashion-gen: The generative fashion dataset and challenge,” arXiv preprint arXiv:1806.08317, 2018.

[28] S. Guo et al., “The imaterialist fashion attribute dataset,” Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0-0, 2019.

[29] F. Carlsson, P. Eisen, F. Rekathati, and M. Sahlgren, “Cross-lingual and Multilingual CLIP,” Proceedings of the Language Resources and Evaluation Conference, pp. 6848-6854, 2022.

[30] OpenAI, “GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, 2023.